Expert Systems With Applications 62 (2016) 190–201

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Natural gas consumption forecasting for anomaly detection

Lorenzo Baldacci[a], Matteo Golfarelli[a,*], Davide Lombardi[b,1], Franco Sami[b,1]

[a] DISI - University of Bologna, Via Sacchi, 3, 47521 Cesena, Italy
[b] HERA Group - Via Grigioni 19, 47122 Forlì, Italy

ARTICLE INFO

Article history:
Received 27 January 2016
Revised 8 June 2016
Accepted 8 June 2016
Available online 16 June 2016

Keywords:
Anomaly detection
Linear regression
Local regression
Gas flow forecast

ABSTRACT

Natural gas consumption forecasting is critical for many gas supplier companies tasks - e.g. gas procurement optimization, pipe network monitoring, management and security. This paper presents the joint work we carried out with HERA S.p.A., Italian gas provider leader, which goal is to forecast gas consumption for a given gas network as well as detecting anomalous gas flows according to historic data so to facilitate the monitoring and security processes in their central control room.

Historic network conditions are sampled every 15 min, each sample is composed by a gas flow, an outside temperature, and the timestamp the sample was recorded. Descriptive analyses were carried out using historic data in a village and a small city, then two forecasting techniques were defined, one based on a nearest neighbor approach, one employing local regression analysis.

Experimental results show that the historical data collected and stored can be used to reliably forecast gas consumption. A quantitative and qualitative comparison of the two methods is discussed in details so to highlight strengths and weaknesses. Moreover, due to the peculiarity of the domain, we worked with domain subject-matter experts to understand the capability of the methods in detecting anomalous gas consumption.

Our results clearly show our forecasting techniques effectively support control room operators in identifying anomalous consumption. Providing a forecasting functionality is the first relevant step towards creating a full expert system that makes it easier for advanced operators to interpret the gas network behavior and that suggests the less-skilled ones the correct reactions to be taken upon the occurrence of anomalous events.

1. Introduction

Natural gas is the cleanest energy source over the fossil fuels. Abundant gas resources, increasing investments, and long-term contracts contribute, year after year, to lower gas price, improve infrastructures, and increase its usage in residential, industrial environments as well as transportation. Due to this mix of reasons natural gas consumption has grown significantly over the past years, moreover having seen the European Commission efforts on lowering carbon energies this trend is forecasted even to raise.

Worldwide, natural gas, is mainly used for residential, commercial and industrial environments; also electric energy generation significantly contributes to gas consumption whilst transportation still plays a secondary role.

In Italy, gas distribution to users is managed by regional gas suppliers who are responsible for gas procurement, distribution and pipe network management. They faces many challenges in their business like forecasting gas demands for optimizing procurement, pipe network monitoring for maintenance and security reasons, gas meters reading for user profiling, customer segmentation, fraud prevention, etc. Among these challenges, those linked to security and service delivery are obviously of main concern. With this in mind, monitoring gas flow and pressure within the pipe network represent a key activity for many security and service related control workflows.

This research and experimental work is an outcome of a joint collaboration between the University of Bologna and Hera S.p.A., which is a regional company, manly operating in Emilia–Romagna, for gas, water, electricity and drainage services. Hera aims at improving its gas services because they are, among the others, those

* Corresponding author.
E-mail addresses: lorenzo.baldacci@unibo.it (L. Baldacci), matteo.golfarelli@unibo.it (M. Golfarelli), davide.lombardi@gruppohera.it (D. Lombardi), franco.sami@gruppohera.it (F. Sami).
[1] Honorary author

http://dx.doi.org/10.1016/j.eswa.2016.06.013

mainly affected by security law regulations and needs for proactive actions.

In particular Hera faces a challenge in managing their gas network: understanding anomalous gas flows. When gas flow shows anomalies, it causes leaps in pressure. There are two kinds of pressure leap: negative leaps which could cause end user meters to be locked due to security devices installed on them; and positive leaps which could cause high stress to network components and, when extreme, cause bursts.

For example a negative spike of pressure causes meters to lock and it can only be unlocked via manual intervention. This is due to a security policy regulating the unlocking process, the gas engineer is allowed to proceed only with the customer's supervision in order to avoid gas leaks inside its building. As per positive pressure spikes, Hera historically experienced a pipe burst due to high pressure in 2006. The leakage originated an explosion and a building collapsed causing casualties.

Over the past 5 years, Hera has started a journey for improving his monitoring capabilities over its gas pipe network. Along this journey, Hera, have installed an advanced sensor network where gas pressure and flow are read on any gateway (i.e., entry point linking the Italian national gas network and the local networks where end users are connected to). Sensors communicate with the central data bank providing real-time meter's readings.

The sensor network has now collected a huge quantity of readings. Hera would like to leverage such information to create gas usage model that allows to predict (1) the gas consumption with a high forecasting resolution and a short forecasting horizon, (2) the consumption window that marks normal consumption variations from anomalous peak. Such forecasting functionality can be immediately exploited by the control room users to judge the overall network status by comparing the predicted values against the values read by the sensor network. This comparison facilitates the detection of anomalous gas flows undergoing in the pipe network and can be used to raise automatic warnings.

In this work we present two techniques to model the natural gas usage in terms of predicted consumption and normal consumption window. In particular the first technique is based on a stochastic model whilst the second is based on a local regression analysis. More in details the original contributions of the paper are:

- An in-depth analysis of the Natural Gas Consumption domain based on the data made available from Hera (see Section 2).
- The definition of two forecasting techniques the desiderata and constraints provided by the Hera domain experts (see Section 3.1 and Section 3.2).
- A large set of tests aimed at evaluating the effectiveness and efficiency of the two approaches. The comparison between the results makes it possible to better understand strength and weakness of the two methods and let the user to achieve a better understanding of gas domain (see Section 4).

The paper organization is completed by an analysis of the related literature (see Section 5) and by the conclusions that also outline our future and current researches (see Section 6).

We emphasize that the final goal of the HERA project is to develop more advanced functionalities on top of our forecasting techniques. The complete system will show both decision support system and expert system features and will be aimed at supporting the control room operator to understand the gas network behavior and to make better and faster decisions. One of the problem faced in control rooms is the limited experience of some of the operators. For this reason, HERA would (1) provide the operators with a set of tools for easily interpreting the network behavior; (2) code in an expert system the knowledge of the most skilled operators to enable the correct reactions to be taken upon the occurrence of anomalous events. In particular, interpreting the consumption patterns and the drift from the forecasted values require some intelligence. For example a smooth, progressive drift from the forecasted values could have a different interpretation with respect to a sharp peek of consumption. Peak shapes may also be related to the season, weather conditions and the network specificities (e.g. number of inhabitants, presence of industries).

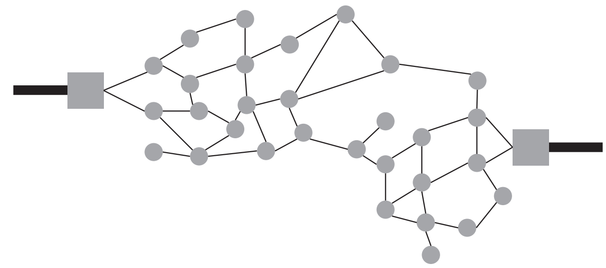

Fig. 1. A typical local gas network with several customers (circles) and two RMPs (squares).

2. Natural gas consumption: data and domain analysis

In this section we report the outcomes of the analysis we carried on the natural gas consumption domain. This rather critical step is aimed at acquiring a deep understanding of the domain specific features and it lays down the basis on which our forecasting models will be built on. Indeed, it highlights data quality issues that would potentially invalidate the prediction model on the one hand and it points out the domain's dependent and independent variables on the other.

The data are provided by HERA S.p.A., one of the italian's leaders in energy and water infrastructure management and service provider. In term of natural gas distribution, HERA mainly operates in the region of Emilia–Romagna, it manages around 14,000 Km of gas network servicing more than a million customers. The overall gas network is partitioned in several local networks typically servicing a well-defined geographical area corresponding to a town or part of a larger city. Each local network is supplied through one or more Regulation and Measurement Plant (RMP) that are the borders between the national provider network and the HERA's network. The gas flow for each local network is computed as the sum of the gas flows passing through the RMPs supplying such network. Fig. 1 shows a typical local gas network supplied thorough two RMPs (squares).

The data used in this work are related to gas consumption volumes in two different networks: *Network 1* services Voltana, a village of 2,200 inhabitants mainly including residential utilities; *Network 2* services Lugo, a small city of 33,000 inhabitants including, besides residential buildings, some commercials and offices. Both the areas shares a the same temperate climate with hot summers (with highest temperature over 30 °C degrees) and cold winters (with lowest temperatures below zero).

The dataset comprises about 35,000 observations per year per network over a period of about 3 years in total, from July 2010 to September 2013. Besides the historical values HERA provides the local temperature forecast for the next 12 h.

Definition 1 (Observations Dataset)**.** Given a natural gas network, an Observations Dataset $S = \{s_1, s_2, \ldots, s_n\}$ is a sequence of n historical samples, where each sample $s_i = (g, t, d)$ is characterized by the gas flow g expressed in Stm^3/h, the local outside temperature t in °C degrees, and a time stamp d.

Samples are collected by the HERA's SCADA system every 15 min that is the finest time precision in our project (in each day

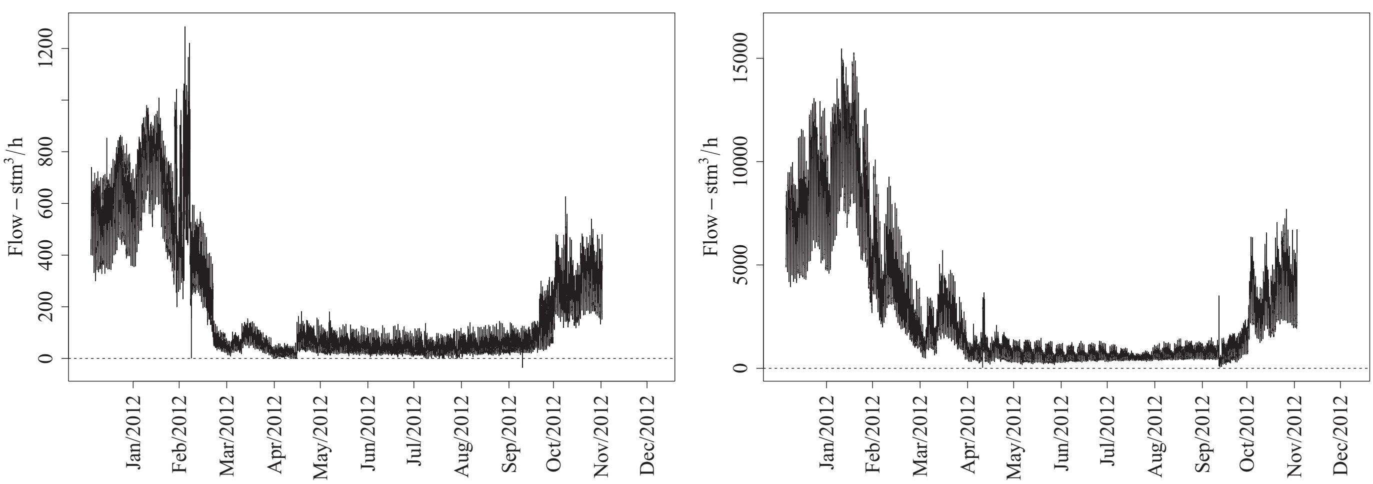

Fig. 2. Gas Flow - Yearly profiles for Network 1 (left) and Network 2 (right).

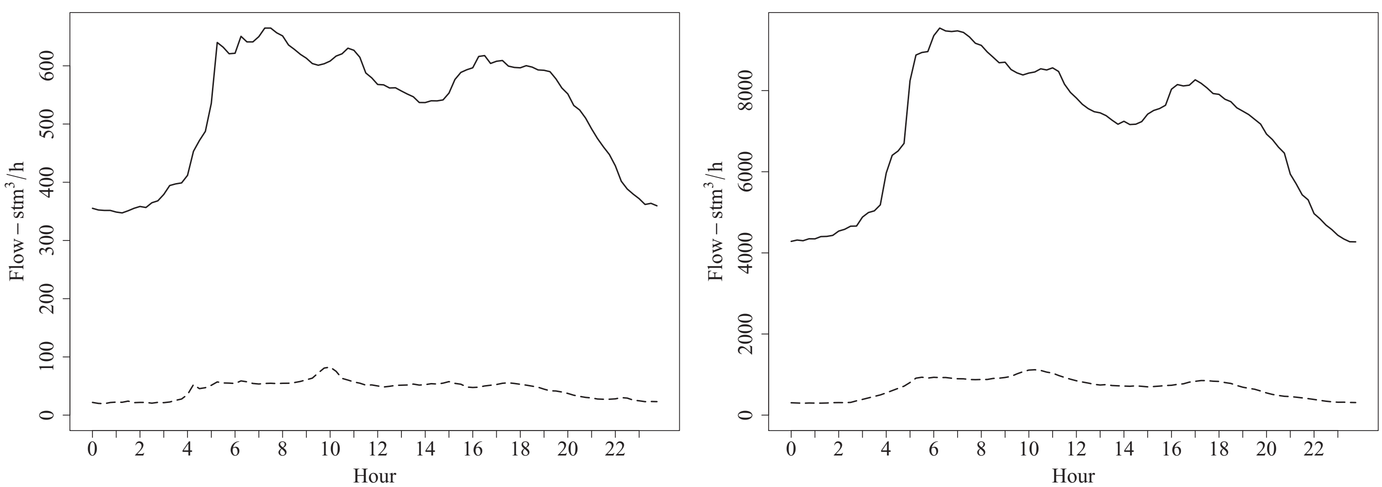

Fig. 3. Gas Flow - Daily profiles for Network 1 (left) and Network 2 (right). In each figure solid line represents the winter profile, while the dashed one represents the summer profile.

96 *time-window* are defined). The time stamp allows to determine the season, the day of the week and the type of the day the sample has been collected in. More formally, given a samples s, it is $DType(s.d) \in$ {*'Non-working', 'Working'*}, $Seas(s.d) \in$ {*Spring, Summer, Autumn, Winter*}, $Month(s.d) \in \{January, \ldots, December\}$, $Day(s.d) \in$ {*Monday,..., 'Sunday'*}, and $TWin(s.d) \in \{1, .., 96\}$.

Fig. 2 shows the gas flow for the raw historical data: as expected a year/season-based consumption pattern is apparent. Winter months show gas flows that are up to an order of magnitude higher than the summer ones mainly due to the heating consumption. The figure also shows a couple of anomalous peaks (i.e. Network 1 - March, Network 2 - December) that are due to missing values in the data sets. Besides the main pattern, Fig. 2 is characterized by a very short-period fluctuation of the gas flow that results in a stripe instead of a thin line. This is better explained when moving the time series analysis from year to day. Fig. 3 makes apparent a daily customer behavioral pattern characterized by morning and evening peaks as well as the lower request during night time. Peaks and valleys are much more evident in winter due to the heating consumption.

The daily profile changes depending on the day of the week too. Fig. 4 compares *daily profiles* obtained by averaging gas consumptions at a given hour and in a given day of the week. The graphs are reported for Network 2, Network 1 shows a similar behavior. These figures emphasize two different behavioral patterns for different types of day. In particular working days share the same pattern, while during Saturday and Sunday gas flow is always lower. The differences are more evident in summer as in winter heating plays a major role compared to human behavior and therefore it partially hides this pattern. In order to get a better understanding on the relationships between daily profiles in different days of the week, we computed the similarity between them according to the following formula.

$$Sim(V_{x',y'}, V_{x'',y''}) = \frac{1}{96}\sum_{i=1}^{96}\frac{|v^i_{x',y'} - v^i_{x'',y''}|}{max(v^i_{x',y'}, v^i_{x'',y''})}$$

where $V_{x,\,y}$ is the set of the average gas consumption samples in S for each time-window in a given day of the week x and in a specific season y with $i \in [1, .., 96]$. More formally $v^i_{x,y} = avg(s_i.g)|s \in S \wedge Day(s_i.d) = x \wedge Seas(s_i.d) = y$. Fig. 5 shows such similarity matrix for Network 2. The matrices confirm that working days show a similar usage pattern. Since the working days behave similarly in this work we will group days from Monday to Friday and we will consider only two day-types: working and non-working.

So far we have sliced the gas consumption time series by, time of day and day of the week. Fig. 6 shows a sample of slices plotted by gas consumption and temperature (keeping season, time and day type constant). The graphs emphasize that, keeping fixed the other variables, there is a strong linear dependency between

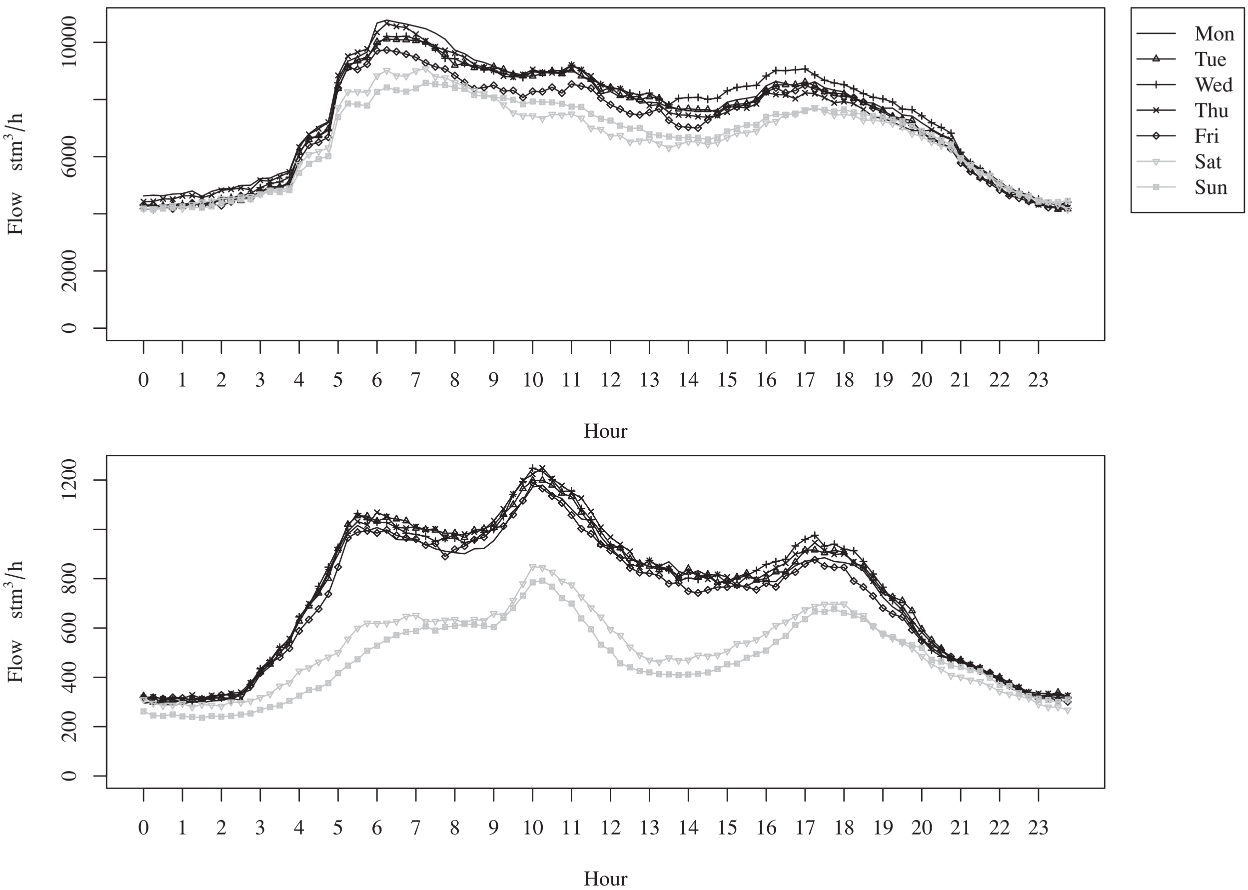

Fig. 4. Network 2 Gas Flow - Daily profiles in different days of the week distinguishing between winter (Top) and summer (Bottom) months.

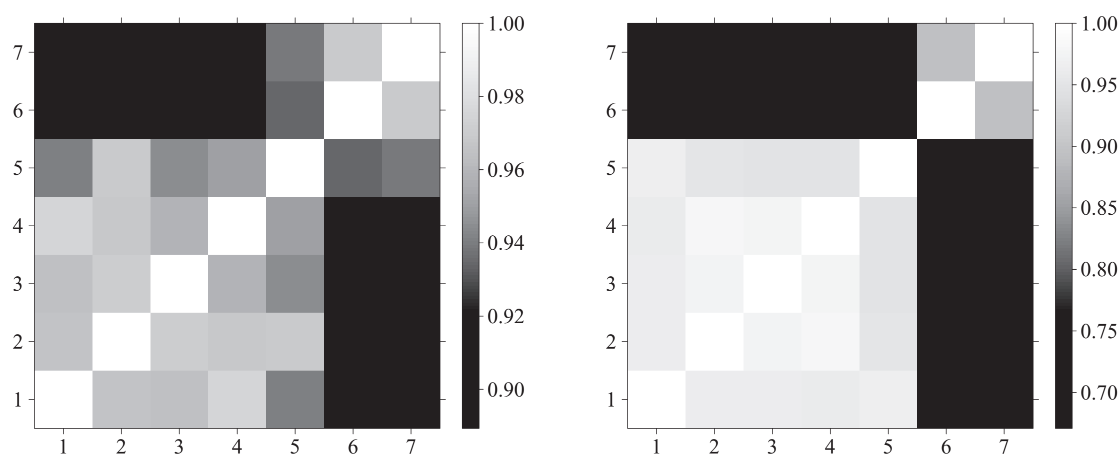

Fig. 5. Similarity matrices for gas consumption in different days of the week for Network 2 in winter (left) and summer (right). Days range from 1 = *Monday* to 7 = *Sunday*.

the temperature and the gas consumption in winter and summer.[2] Conversely spring and autumn require further examination. The autumn and spring pattern can be explained considering that in these seasons temperature can easily change of many degrees during the same day and cold periods can be followed by warm ones. Domain experts reported us that in autumn, if a cold period induced inhabitants to turn the heaters on, they will not be willing to turn it off later on evan if temperatures increase and thus a typical winter consumption will take place. The opposite behavior could take place in spring. In other words, the human behavior (that we will call *temperature recall*) affects the linear dependency between temperature and gas flow that, in these seasons is further faded by the temperature variability.

Finally, a further discussion must be devoted to the temperature forecasts $\bar{t}$ that in our case are provided by the Regional Agency for the Environment (ARPA) exploiting the COSMO model (Cosmo Consortium, 2015). As said before, every 12 h ARPA makes available the temperature forecasts for the following 48 h. Since such forecasts will be an input for our models, an error Δ_t between the forecasted temperature $\bar{t}$ and the observed temperature t may impact on the model's accuracy itself. Fig. 7 shows the

[2] During summer gas consumption does not change for increasing temperature. This pattern can be modeled by a constant linear function.

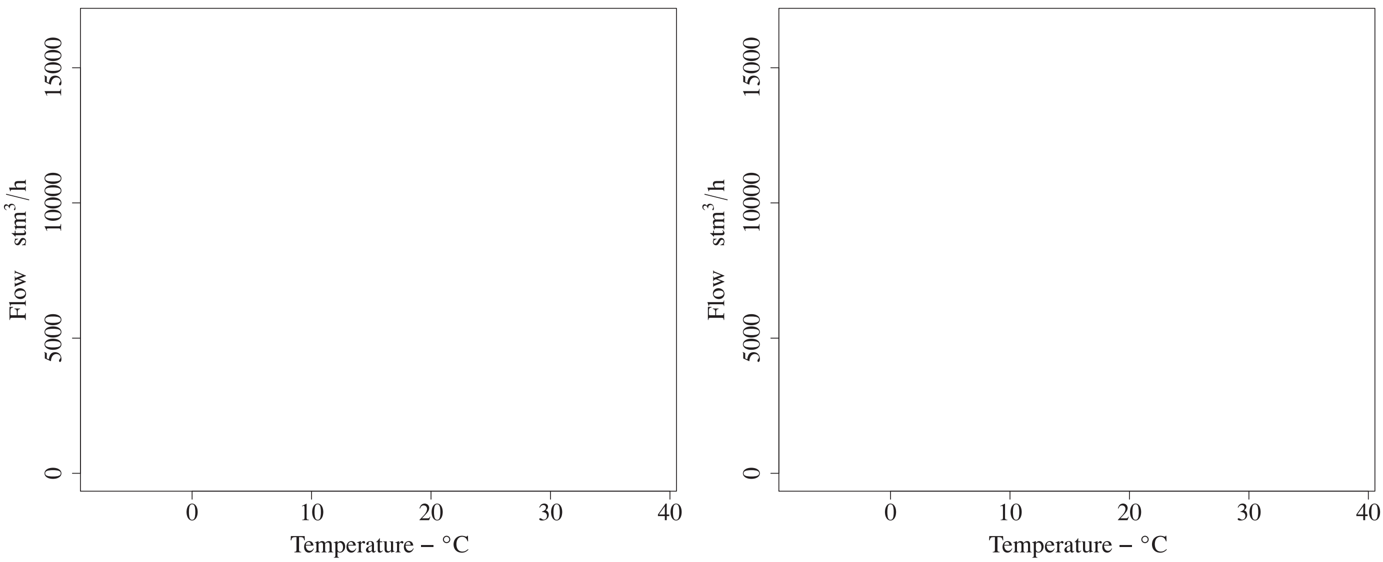

Fig. 6. Gas Flow - Temperature correlation for Network 2 at 6pm in winter (left) and summer (right).

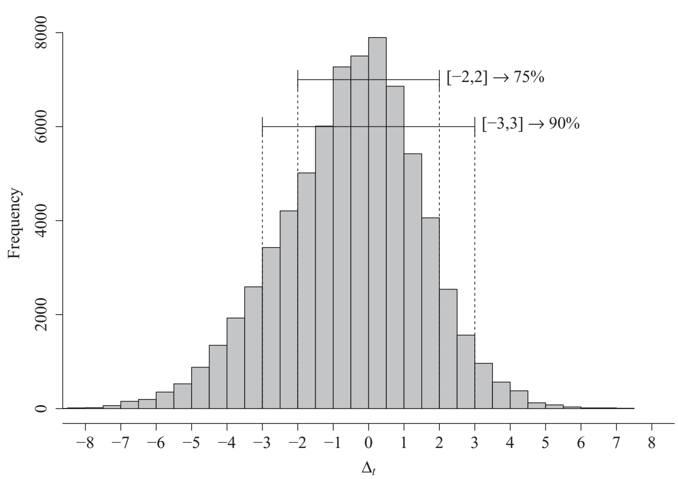

Fig. 7. Network 2 - Histogram of the ARPA temperature forecasting errors in Celsius degrees.

histogram reporting the temperature forecasting errors calculated as the difference between forecasted and observed temperature. Noticeably, the figure shows that 75% of the samples lie within the range $[-2, 2]^{\circ}C$, the ratio raises to 90% if we consider the $[-3, 3]^{\circ}C$ range. In Section 4 we will evaluate the impact of such errors on our models.

3. Gas consumption forecasting

Based on the analysis carried out in the previous section, we propose two different algorithms for the Gas Consumption Forecasting Problem that can be defined as follow:

Definition 2 (Forecast Problem)**.** Given an Observation Dataset $S = \{s_1, \ldots, s_n\}$ storing the historical gas consumption for a specific natural gas network and knowing the forecast temperature $\bar{t}$ at time stamp d our goals are (a) to forecast the gas consumption $\bar{g}$ at time d; (b) to estimate the interval $[\bar{g}_{min}, \bar{g}_{max}]$ the forecasted consumption will fall in with probability *Prob*.

Providing an estimate of consumption interval is aimed at distinguishing the normal consumptions from the anomalous ones: by setting one or more probabilities (e.g. 75%, 95%, 98%) the users will be able to define increasing levels of attention that will be associated to different actions to be undertaken. One of the user desiderata is to minimize the complexity of the tuning and maintenance phases that deeply impact on the system performances and is in most of cases carried out by non-expert staff with a limited sensibility to parameters and algorithms. For this reason, in Section 3.1 we propose a *Lazy Forecast* approach that does not rely on a forecast model and thus does not require any periodic maintenance activity. Conversely, the one discussed in Section 3.2, relies on a forecast model that must be periodically trained in order to consider the latest change in network and inhabitants behaviors.

3.1. A lazy forecast approach

This approach, hearafter referred to as *LF*, belongs to the *nearest-neighbor* family. It exploits the historical samples with similar features to compute the forecasted consumption defined as the median of subset $H_{d,\bar{t}}$ of selected samples.

$$H_{d,\bar{t}} = \left\{ s \in S | \bigwedge \begin{array}{l} \bar{t} - \Delta_t \leq s.t \leq \bar{t} + \Delta_t \\ TWin(d) - \Delta_{tw} \leq TWin(s.d) \leq TWin(d) + \Delta_{tw} \\ Seas(s.d) = Seas(d) \\ DType(s.d) = DType(d) \end{array} \right\}$$

Since the analysis of the domain evidenced different pattern in different seasons and different day types we select only samples that share this features with timestamp to be forecasted. Conversely, we allow a range of variability for time ($\Delta_{tw} = 15'$) and temperature ($\Delta_t = 2^{\circ}C$) variables since consumption changes smoothly for small variation of d and $\bar{t}$. This choice is also aimed at computing the median on a set of samples $H_{d,\bar{t}}$ whose cardinality is statistically relevant. Furthermore, since temperature forecasting is not immune by small errors a tolerance interval can be justified. As we will show later in Section 4, when $|H_{d,\bar{t}}|$ is low the forecast correctness tends to decrease. For this reason the lazy forecast approach does not return any forecast when $|H_{d,\bar{t}}| < mincard$.

$$\bar{g} = LazyForecast(d, \bar{t}) = \left\{ \begin{array}{l} median(h.g \text{ s.t. } h \in H_{d,\bar{t}}) \text{ if } |H_{d,\bar{t}}| > mincard \\ \text{undefined otherwise} \end{array} \right\}$$

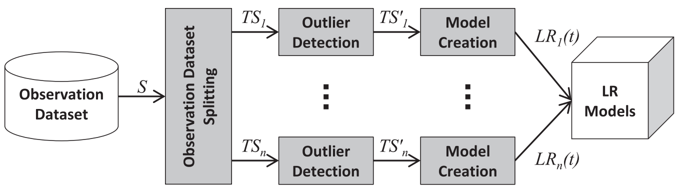

Fig. 8. Steps composing the local regression approach.

The *mincard* value determines a trade-off between the percentage of forecasted values and their correctness; similarly. Inability to forecast is more frequent when anomalous temperatures (too high/low) are forecasted for a specific date and time, in these cases few or none historical samples are present in $H_{d,\bar{t}}$. Higher values for Δ_t and for $\Delta_t w$ enlarge the considered neighbor samples at the risk of considering non-representative ones. We tuned the thresholds by testing on a training dataset different combination of values minimizing the prediction error $|\bar{g} - g|$ while returning a forecast in an acceptable percentage of times.

The consumption intervals are estimated through the percentile of the consumption values in $H_{d,\bar{t}}$. More precisely:

$$\bar{g}_{min} = s.g \in H_{d,\bar{t}} | s.g \texttt{ is } \frac{100 - Prob}{2} - th\ \ percentile$$

A similar definition can be given for $\bar{g}_{max}$.

Apart from setting the thresholds, the lazy approach does not require any learning process and it has been implemented on the RDBMS storing the Observation Dataset through SQL queries.

3.2. A local regression approach

The second approach we propose is based on Local Regression - *LR* (Cleveland & Devlin, 1988; Loader, 2006). The main advantage of this class of techniques is that they enable the fitting of complex data spaces without using global regression functions that are implicitly difficult to tune. To reach this goal, several simple regression functions, each modeling a small part of the data space, are combined to create a complex model.

In our specific case we build on linear regression functions whose explanatory variables are the observed temperature and whose dependent variable is the gas consumption. In order to model the observation dataset *S* we combined 768 regression functions resulting from splitting the data set according the day type (2 values), the season (4 values) and the times of the day based on a 15 min time windows (96 values). According to the analysis discussed in Section 2 such subsets evidence a linear trend and thus can be effectively modeled through a linear regression function.

The training process steps are depicted in Fig. 8. Initially the observation data set is split in several subsets $TS_{dt,\ se,\ tw}$ each related to a specific part of the data space (i.e. a day type *dt*, a season *se* and a time-window *tw*) that will be modeled by a single regression function. For each subset, an outlier detection algorithm is applied in order to remove possible outliers that may twist the model. Finally, the obtained training sets $TS'_{dt,se,tw}$ are used to create the linear models whose union determine the *non-linear* model that covers the whole data space. In the following we will discuss each single step in more details.

3.3. Observation data set splitting

Observation data set splitting is the step that deserves more explanations. Using a such high number of regressive functions ensures both a linear behavior of samples in each subset and it allows a very detailed modeling of the data space. Samples are associated to a unique label as concerns day type and time window while they can be associated to two labels of consecutive seasons. Please note that, differently from many other approaches, we split the samples on a four-seasons base since, although the dependent variable (consumption) is always linearly related to the explanatory one (temperature), observation scattering varies in different seasons. Scattering is particularly evident in spring and autumn that are characterized by wide variation of daily temperatures and the turning on/off of the heating systems that can be anticipated or postponed by house owners depending on the occurrence of warmer or colder periods. Scattering makes spring and autumn forecasts particularly complex and suggests to adopt specific models for them.

In the areas we tested this model the periods which evidenced the highest scattering are $\{March, April\}$ and $\{October, November\}$. We will label the first one as Spring and the second one as Autumn. Overall seasons are defined as follows:

$$Seas(s.d) = \begin{cases} spring & Month(s.d) \in \{March, April\} \\ summer & Month(s.d) \in \{May, June, July, August, September\} \\ autumn & Month(s.d) \in \{October, November\} \\ winter & Month(s.d) \in \{December, January, February\} \end{cases}$$

Although we use the traditional season labels, our seasons do not coincide with the meteorological ones but rather they cluster periods that share the same behavior and a similar scattering of the samples. Obviously such choice may vary according to the specific region the model is applied to.

Splitting seasons based on a sharp temporal-based criterion only is not the best choice since the transition between the behaviors adopted in two consecutive seasons can be actually anticipated or postponed in different years according to the occurrence of warmer or colder temperatures. We explicitly modeled such fuzziness in season transitions by coupling the temporal-based criterion with a temperature-based one.

In the energy forecast applications the temperatures that mainly impact on the consumption are those that determine the turning off/on of the heaters and coolers. Such threshold is also called *base temperature* which acts as reference temperature for those methods that employ the concepts of Heating Degree Days (*HDD*) or Cooling Degree Days (*CDD*).[3] In the literature base temperature splitting is often tackled by means of a sharp threshold upon the daily average temperature (Büyükalaca, Bulut, & Yılmaz, 2001). In the present work we built on this idea for modeling a more complex behavior that we call *temperature recall*. According to the domain experts we worked with householders tend to turn the heating on/off in autumn/spring when a particularly cold/warm period takes place but afterwards they will not turn it off/on if the temperatures increase/drop again. In the areas we tested this model to these temperature turned out to be $t^*_S = 14.5$ $C°$ for turning off, and $t^*_W = 12.0$ $C°$ for turning on, where the temperature considered for a given sample *s* is the average of temperatures observed for the samples belonging to the previous 48 h: $AVG48(s_i.t) = AVG_{j=191}^{j=0} s_{i-j}.t$. Fig. 9 shows a typical example of this behavior: at time d^* an heavy drop of temperatures determined a strong increase of gas consumption due to heaters turning on, but afterwards, even if temperatures rise again over the t^*_w thresholds, consumption remain high and follows a typical winter trend proving that householders kept their heating on.

Based on this behavior we defined a temperature-based season *TSeas*(*s.d*). Season transitions are triggered by the period of the year and the thresholds t^*_S and t^*_W and can be modeled according the following rules:

[3] HDD and CDD are defined as the amount of energy necessary for heating or cooling a building given the weather condition it is exposed to.

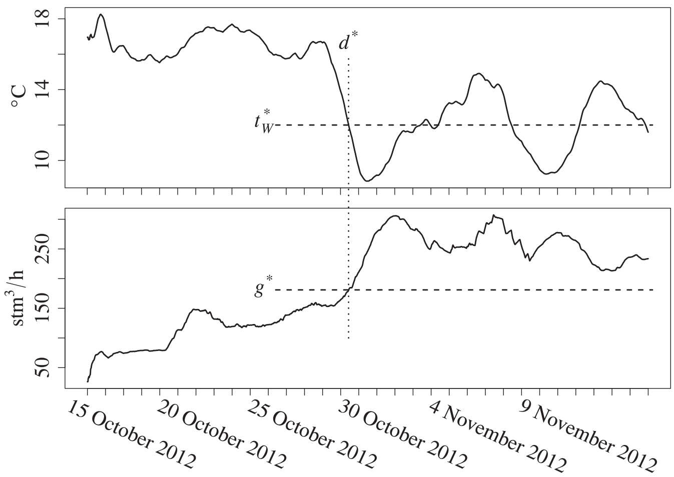

Fig. 9. 48 h moving average temperature (top) and gas consumption (bottom). Temperature drop at d^* triggers a change between a *summer* and a *winter* behavior: gas consumption before and after such date are differently related to the temperatures.

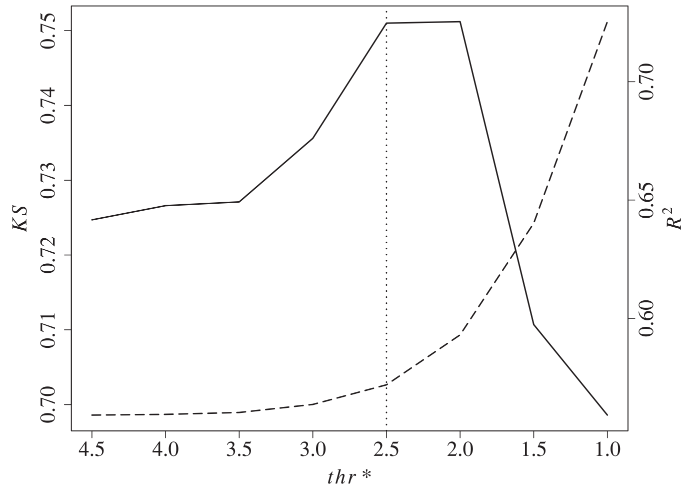

Fig. 10. Kolmogorov-Smirnov values (solid line) and R^2 values (dashed line) varying the outlier's threshold.

$$TSeas(s_i.d) = \begin{cases} spring & TSeas(s_{i-1}.d) = winter \wedge \\ & \wedge Month(s_i.d) \in \{February, March\} \wedge AVG48(s_i.t) > t^*_W \\ summer & TSeas(s_{i-1}.d) = spring \wedge \\ & \wedge Month(s.d) \in \{April, May\} \wedge AVG48(s_i.t) > t^*_S \\ autumn & TSeas(s_{i-1}.d) = summer \wedge \\ & \wedge Month(s.d) \in \{September, October\} \wedge AVG48(s_i.t) < t^*_S \\ winter & TSeas(s_{i-1}.d) = autumn \wedge \\ & \wedge Month(s.d) \in \{November, December\} \wedge AVG48(s_i.t) < t^*_W \\ TSeas(s_{i-i}.d) & otherwise \end{cases}$$

Note that: (a) *TSeas* can change only between two consecutive seasons (i.e. winter to spring, spring to summer, etc.); (2) the rules implement the temperature recall behavior since once we move from autumn to winter due to a cold period, further temperature rises does not move back the *TSeas* to autumn.

It is apparent that meteorological and temperature-based seasons are overlapped and the start/finish of the temperature-based ones is not sharp but may vary over a couple of month. When $Seas(s.d) \neq TSeas(s.d)$, s will be part of two training sets. More formally, the training set *TS* for a day type *dt*, a season *se* and a time-window *tw* is defined as follows:

$$TS_{dt,se,tw} = \{s \in S | DType(s.d) = dt \wedge (Seas(s.d) = se \\ \vee TSeas(s.d) = se) \wedge TWin(s.d) = tw\}$$

This features determines each single regression line to flexibly model the different meteorological period that characterize season transitions.

3.4. Outlier detection

In the local regression approach, for each part of the data space determined by *dt*, *se* and *tw*, the gas consumption forecast is computed through a linear regression model $LR_{dt,se,tw}(t) = \alpha_0 + \alpha_1 t$. We employ the classic least squares method (Montgomery, Peck, & Vining, 2012) for fitting the linear model to the training data.

In order to obtain a more accurate forecast, before creating the final model we carry out an outlier detection step aimed at identifying and dropping from training set $TS_{dt,\ se,\ tw}$ outliers that would twist the model (Rousseeuw & Leroy, 2005). The outlier detection for a subset $TS_{dt,\ se,\ tw}$ is achieved by fitting a linear model on it too, then by analyzing the obtained residuals. More formally, let the following be the equation defining the linear fitting problem:

$$s_i.g = \alpha_0 + \alpha_1 s_i.t + \epsilon_i$$

where $s_i \in TS_{dt,\ se,\ tw}$. α_0 and α_1 are the regression model's coefficients, whilst ϵ_i model residual (i.e. the difference between the observed gas flow $s_i.g$ and its fitted value $LR_{dt,\ se,\ tw}(s_i.t)$).

For each model *LR*, residual analysis returns residual distribution that are supposed to be normal distributed, having mean 0, and standard deviation $\sigma_{dt,\ se,\ tw}$ (i.e. $\mathcal{N}(0, \sigma^2)$). In order to determine a unique threshold thr^* we must studentize (Rousseeuw & Leroy, 2005) the residuals distributions, this will make them comparable across different *LR* models.

An outlier is defined as an observation having its studentized residual $\tilde{\epsilon}_i$ higher than thr^*, such outliers are excluded from $TS_{dt,\ se,\ tw}$ obtaining the training set $TS'_{dt,se,tw}$.

$$TS'_{dt,se,tw} = \{\forall s_i \in TS_{dt,se,tw} | abs(\tilde{\epsilon}_i) \leq thr^*\}$$

The outlier detection comes along with the problem of setting the threshold thr^*. This determines a trade-off between two goals: (1) effectively cutting-off outlier samples; (2) avoid non-outiler samples lying in the distribution tails to be cut-off, thus loosing relevant information.

How well a threshold match such goals can be measured through two tests:

- KolmogorovSmirnov test (Lilliefors, 1967) (*KS test*): measures the normality of residuals. Choosing the threshold thr^* impacts the probability distribution of residuals and, on the one hand a lenient threshold would not cut off outliers resulting the residuals' distribution to significantly differ from the normal, on the other hand, even being too strict would cause the residuals' distribution tails to be cut off, still causing the residuals' distribution to differ from normal.
- Coefficient of determination (Montgomery et al., 2012) (R^2): measures how well the training set is described by the model. The higher R^2 is, the better the training set is modelled. In our context, R^2 coefficient is used to quantify how well *TS* is described by the linear model and then to measure how much TS' differ from *TS*. As per the R^2 test, lenient values for t^*_O would cause little or no difference between *TS* and TS' as just few outliers would be detected, whilst strict values for t^*_O would cause TS' to be better described by the linear model due to the fact that observations which do not lie close to the regression line would be excluded from training. With this is mind we would like similar R^2 for either *TS* and TS' as to avoid significant loss of information caused by the outlier detection step.

Fig. 10 shows *KS* and R^2 for decreasing values of thr^* on *TS* and on TS'. As per *KS*-test (a) when a strict threshold (i.e. close to 1) is applied distribution differs from normal since since distribution tails are cut-off. Similarly, when a too lenient threshold is chosen (i.e. values close to 4.5) are not cut-off determining a non-normal distribution. As per R^2 strict thresholds determine larger differences in the models for TS and TS' thus emphasizing an information loss.

Based on the outcomes in Fig. 10 the best threshold value is 2.5. Is worth noting that Rousseeuw and Leroy (2005) proposed an empirical value of 2.5σ as threshold for outlier detection endorsing therefore our choice. Setting $thr^* = 2.5\sigma$ in the TS' definition resulted in 1.46% of observations contained in TS to be marked as outliers (i.e. 309 out of 21,150).

3.5. LR-based model in action

The outlier detection step returns, for each training set $TS_{dt,\,se,\,tw}$, a cleaned dataset $TS'_{dt,se,tw}$ which is used as training set for obtaining the final local model $LR_{dt,\,se,\,tw}$. Such model is simply obtained by refitting a new regression model to the training set $TS'_{dt,se,tw}$. Given the set of local models LR, the gas consumption forecast is obtained as follows:

$$\bar{g} = LocalForecast(d, \bar{t}) = LR_{DType(d),TSeas(d),TWin(d)}(\bar{t})$$

As to $[\bar{g}_{min}$ and $\bar{g}_{max}]$, they are computed as the prediction interval for the given probability $Prob$:

$$[\bar{g}_{min}, \bar{g}_{max}] = \bar{g} \pm t^*_{Prob}\sigma\sqrt{1 + \frac{1}{n} + \frac{n(\bar{t} - \hat{t})^2}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2}}$$

where t^*_{Prob} is the t-percentile from a t-distribution. $\hat{t}$ and n are the mean temperature and the number of the observations in $TS'_{dt,se,tw}$ respectively. Finally, t_i is the i^{th} observation within $TS'_{dt,se,tw}$.

4. Results and comparison

In this section we discuss the experimental results obtained with the two proposed methods on the data available for Network 1 and Network 2. The networks and their features have been deeply discussed in Section 2. Each Observation dataset S includes two years of samples (i.e. 1^{st}*June* 2011 to 31^{st}*May* 2012); that are split in two subsets of one year each. One of them has been used to train the models while the other one has been used as test set.

The system accuracy has been evaluated in terms of:

- *Forecasting Error* - $Err = AVG_{s\in S'}|s.g - \bar{g}|$
- *Relative Forecasting Error* - $RelErr = AVG_{s\in S'}\frac{|s.g-\bar{g}|}{s.g}$
- *Relative Interval Width* - $RelIntWidth = AVG_{s\in S'}\frac{\bar{g}_{max}-\bar{g}_{min}}{s.g}$
- $Anomaly\% = \frac{1}{|S'|}\sum_{s\in S'}\begin{cases}1 & (s.g > \bar{g}_{max}) \vee (s.g < \bar{g}_{min})\\ 0 & \text{otherwise}\end{cases}$

where $s.g$ is the measured gas flow for a sample s in the test set S'; $\bar{g}$, $\bar{g}_{max}$ and $\bar{g}_{min}$ are the forecasted values considering the forecasted temperature available at time $s.d$.

We initially recall that the Lazy Forecast approach returns a consumption forecast only when the cardinality of $|H_{d,\bar{t}}| > mincard$. In Fig. 11 we show the trade-off between forecast accuracy and percentage of returned forecasts (*Forecast%*) for $N2$. It is apparent that while the relative error curve tends to flatten, the forecast percentage decrease almost linearly for increasing *mincard* values. Based on this result we set $mincard = 10$ for all the tests in the following, since it represent a good compromise between the correctness of the results and an acceptable percentage of times such results are provided.

Table 1 and Table 2 reports the accuracy metrics for LF and LR respectively. The main considerations are reported below:

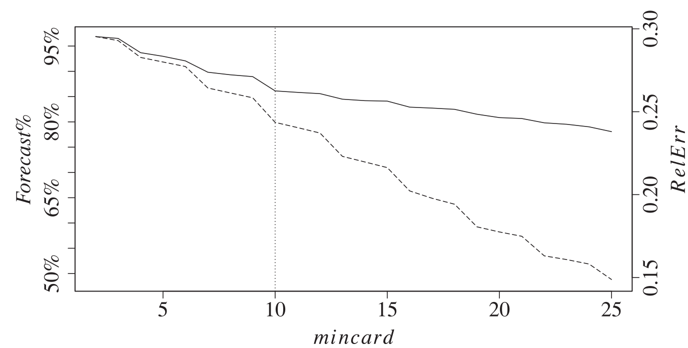

Fig. 11. Lazy Forecast Approach: relative error (solid line) and percentage of forecast (dashed line) varying the minimum number of historical samples needed to provide a forecast.

Table 1
Prediction accuracy indices for *LF*.

	g_i	*Err*	*RelErr*	*RelIntWidth*	*Anomaly%*	*Forecast%*
N1 Spring	279	137	0.51	2.47	16.52%	81.78%
N1 Summer	62	14	0.21	0.95	10.70%	83.88%
N1 Autumn	203	81	0.41	1.88	14.67%	77.34%
N1 Winter	551	45	0.08	0.66	2.92%	99.75%
N1 Overall	284	60	0.26	1.23	9.77%	86.35%
N2 Spring	3,679	1,253	0.48	1.23	37.17%	76.91%
N2 Summer	703	115	0.17	1.00	11.82%	82.67%
N2 Autumn	2,416	863	0.39	1.68	15.19%	76.09%
N2 Winter	7,182	844	0.11	0.53	4.50%	92.95%
N2 Overall	3,090	609	0.23	1.00	13.89%	85.21%

Table 2
Prediction accuracy indices for *LR*.

	g_i	*Err*	*RelErr*	*RelIntWidth*	*Anomaly%*
N1 Spring	279	50	0.28	1.55	2.60
N1 Summer	62	12	0.17	1.04	6.90
N1 Autumn	203	61	0.40	2.00	4.27
N1 Winter	551	57	0.10	0.58	4.09
N1 Overall	284	43	0.21	1.24	4.68
N2 Spring	3,679	572	0.23	1.09	4.99
N2 Summer	703	260	0.40	1.65	2.50
N2 Autumn	2,416	569	0.30	1.35	6.71
N2 Winter	7,182	915	0.12	0.55	6.48
N2 Overall	3,090	523	0.28	1.24	4.60

- Overall LR overcomes LF as to the accuracy of predictions. More in details, LR relative errors are by far smaller than the LF's ones in Spring and Autumn due to the more sophisticated technique used to handle these periods.
- Conversely, LF overcomes LR in summer when gas consumption is very low (see g_i and Err values). This happens because LF, when creating the model, uses smaller datasets containing homogeneous observations by temperature whilst LR creates its regressor using bigger datasets containing wider ranges of temperature that depicts a more generic and robust rule that may be less precise for small consumption. Please note that although summer errors maybe relevant in relative terms are limited in absolute ones.
- LR's robustness comes into play when considering anomalies. LR overcomes LF in detecting the correct percentage of anomalies (i.e. 5%). This result is due to the capability of sizing the width of the prediction interval (see *RelIntWidth*) in accordance with the scattering of the historical samples. This outcome can be appreciated comparing the corresponding values of *RelErr* and *RelIntWidth*. We also note that the two methods agree in labeling anomalies in the 79.2% of cases (i.e. the 79.2% of anomalies found by LR are considered anomalies by LF too).

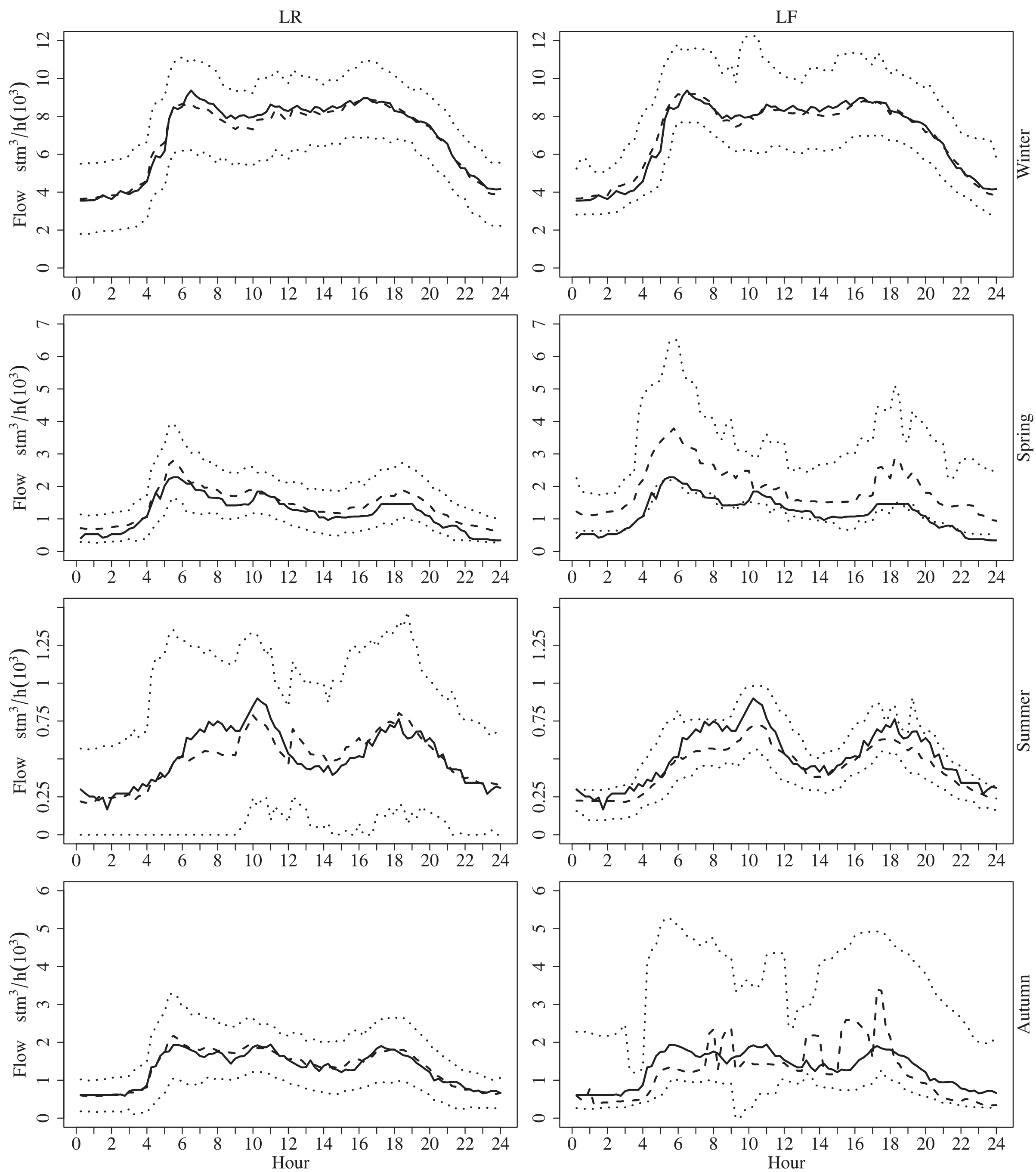

Fig. 12. Hourly prediction graphs for *LR* and *LF* in different seasons. Solid lines represent the actual gas consumptions, dashed lines the forecasted ones. Dotted lines show the intervals the forecasted consumption should fall in with 95% of probability.

- The percentage of times *LF* returns a prediction is lower in spring and autumn that are shorter periods with an higher temperature range. For these reasons it is more difficult to find *mincard* historical samples in the neighborhood of each timestamp *d*.

To better appreciate the behavior of the two techniques under different conditions Fig. 12 reports a detailed output of the two methods in different seasons. This figure enables a qualitative appreciation of the methods, and it shows how the gas network control room operators would benefit by employing a gas forecast tool. It is apparent that while forecasts are, in most cases, very close to the actual values, prediction intervals width changes a lot across seasons and is smoother for *LR*.

To close this section we report some considerations about the specific domain emerging from the domain analysis and tests.

- Forecasting accuracy is strictly related to the cardinality of the sample sets used for training the models. For example, in the climatic area where the tests have been carried out such cardinality varies with seasons. Forecasting is particularly difficult in autumn and spring that are characterized by high temperature variations and by shorter length that determine a lower cardinality of the sample sets $H_{d,\bar{t}}$ and $TS'_{dt,se,tw}$. This negatively impact on both the accuracy of the forecasted consumption and on the interval width.
- High-resolution forecasting problems (i.e. with forecasting frequency higher than 1 per day) are particularly difficult since

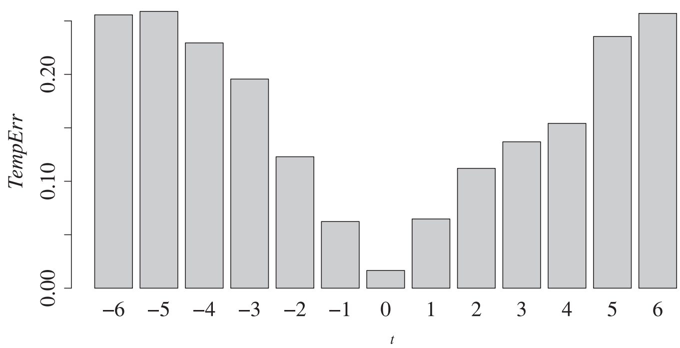

Fig. 13. Portion of gas consumption error due the temperature forecasting error for different temperature errors.

noise and errors, as well as spot effects of uncovered variables, are not eliminated by averaging over time. For this reason, a semantic definition of anomalies cannot ignore the persistence of the outlier values.

4.1. Impact of the temperature forecast error

Accuracy of the two methods depends on the accuracy of the forecasted temperature. Temperature forecasting requires a separated model that is out the scope of the paper. In our framework forecasted temperatures are provided by the Regional Agency for the Environment that exploits the COSMO model (Cosmo Consortium, 2015). Nonetheless, it is important to understand how much such errors impact the accuracy of the gas flow forecast. Fig. 13 shows the fraction of relative error due to the temperature forecast error.

$$TempErr_{\Delta_t} = AVG_{s\in S'}\left|\frac{|s.g-\bar{g}|-|s.g-\hat{g}|}{s.g}\right|$$

where S' is the set of samples with a given Δ_t, $\bar{g}$ and $\hat{g}$ are the gas consumption forecasts computed exploiting the forecasted temperature and the correct temperature respectively.

As expected the larger Δ_t, the larger the fraction of the overall error imputable to the temperature errors. Although the temperature forecasting errors can determine up to the 25% of the overall error, considering the temperature error distribution in Fig. 7 such percentage is 8.8% in the average.

4.2. Efficiency

As per the efficiency, we carried out all tests on a desktop PC dual core (3 GHz, 4GB RAM, Windows 7-64 b it); both methods were implemented in R. Execution times are considerably different: the LF approach does not require any training, while each forecasting requires 0.33 s since a query on the DBMS must be issued. Conversely, the Local Regression approach is much faster in returning each prediction (2.5 ms.) once the model has been created (22.4 s on a one year Observation Dataset). Although computation times could be easily reduced scaling up the hardware resources or carrying out further code optimizations (e.g. pre-loading into the RAM the Observation Dataset in the Lazy Forecast technique), it is apparent that time efficiency is irrelevant with respect to prediction accuracy. In a real domain, such as the one presented, the SCADA system receives the temperature forecasts every 12 h, it has a sampling frequency of the gas flow in each networks of 15' and around 100 gas networks (see Section 2). This makes necessary the computation of 4,800 forecasts every 12 h, which takes 26.4 min with Lazy Forecast and 12 s with Local Regression.

5. Related literature

Prediction of natural gas consumption has been subject of research for decades. The first published studies on this matter started in the middle of the 20th century. Hubbert (1949) investigated the life circle of fossil fuel fields and presented a characteristic curve for their life circle. Such curve would be later named *Hubbert Curve* and used in many paper as forecasting tool for fossil fuel production and consumption. Even in the field of natural gas forecast the *Hubber Curve* has been widely employed even thought it arouse contrasting opinions (Cavallo, 2004). Since these firsts works came out, plenty of works have been proposed and, having seen the criticalities of the applications and the never ending evolution of the energy market, the publishing trend is even due to raise.

The models proposed so far can be categorized using many criteria (Soldo, 2012):

A. *Applied area*. It refers at the wideness of the area considered for forecasting. At the highest level, works have tackled the problem of forecasting of gas consumption at the world level (Behrouznia, Saberi, Azadeh, Asadzadeh, & Pazhoheshfar, 2010), other considered the national level (Sánchez-Úbeda & Berzosa, 2007), other at gas distribution pipe network (Gorucu, 2004a), down to the lowest granularity possible studying single building (Chou & Telaga, 2014) and individual customer profiling (Brabec, Konár, Pelikán, & Malý, 2008).
B. *Input data*. It is critical to select the right blend of information needed to perform the forecasts. Depending on the applied area, the forecasting horizon, and the tools a variety of input data can be used to solve the problem. The weather factor plays a major role in gas consumption therefore climatic data has been widely studied: temperature, heating degree-days, cooling degree-days, humidity, wind speed, wind direction, rain and snow precipitation, etc. Plenty of other types of input data have been taken into account so far: economic parameters (e.g. oil price, GDP, natural gas price, etc.), historic gas consumption, historic energy consumption, surveys, and time related parameters (e.g. seasonality, holidays, day of week, data and time, etc.)
C. *Forecasting horizon*. Many different time horizons have been investigated: from few h , days (Brabec et al., 2008), weeks, months (Kizilaslan & Karlik, 2009), up to years (Gorucu, 2004b) and tens of years (Nashawi, Malallah, & Al-Bisharah, 2010).
D. *Forecasting tool*. In the literature many forecasting techniques have been proposed; as previously mentioned predictions models have been based on the *Hubbert Curve* and its evolution, others employed econometric models (Gelo, 2006), statistical models (Aras & Aras, 2004), artificial neural networks (Dombaycı, 2010; Rodger, 2014) and grey models (Ma & Wu, 2009).
E. *Forecasting goal*. Historically, gas consumption forecasts are mainly used by public administrations as well as by all the sectors of the gas industry to plan their activities (Smith, Husein, & Leonard, 1996). In particular, forecasted consumption are crucial for trading purposes: supply contracts typically include a penalty system that is triggered if the buyer requires in the *next* period (typically one day or one week) a supply that is outside the tolerance interval. Therefore local distribution companies require accurate forecasting models. Forecasting energy consumption has recently become one of the major research field in the energy departments since it helps reducing the energy consumption and increase efficiency (Chou & Telaga, 2014; Rodger, 2014). Domotic devices can nowadays automatically control the house climate basing the decisions on users' schedule. Nest, a company acquired by Google, declared that thanks to its automatically learning thermostat, cus-

tomers saved 11.3% of AC-related energy usage without compromising comfort (Nest-Labs, 2015). Learning people behaviour can save a lot of money but system anomalies still raise and spoil the saving effort. The increase of efficiency introduces a further forecasting goal that is anomaly detection. It is showed (Katipamula & Brambley, 2005) that commercial buildings consume from 15% to 30% more energy than necessary due to poorly maintained, degraded, and improperly controlled equipment (Liu, Jiang, Lee, Snowdon, & Bobker, 2011; Zhang, Chen, & Black, 2011). Obviously, within an anomaly detection application it is not sufficient to forecast gas consumption but it is also necessary to define when the load becomes anomalous (i.e. the concept of outlier). Anomaly detection is also an issue for the gas distributors that must ensure the security of their systems.

Our work can be classified, by means of these criteria, as follow: the applied area is the gas distribution network which may range from few thousands users (i.e. village or small town) to hundred of thousands users (i.e. city and its metropolitan area); as per the input data we consider temperature, time of day, type of day (e.g. working day, weekend, and holiday), and season and previous network loads; our forecasting horizon ranges from few hours to a day with a forecasting resolution of 15 min. In term of forecasting tool we study the employment of two different techniques, one based on a statistical model, and one based on a regression model. Finally, our application is tailored for being used in the control room of a gas distribution network where operators tune the networks behavior based on the forecasted loads and monitor the presence of anomaly peaks in order to ensure security.

To the best of our knowledge no other works in the literature addressed this specific combination of criteria and this makes one-to-one comparison with other approaches almost meaningless. For this reason, we discuss strength and weakness of specific features of the local regression approach (since it outperforms the lazy forecast one) and we limit the comparison to solutions that share with our the goal and the forecasting resolution. Our approach is based on regression functions that have shown to obtain the best detection results if compared with other approaches (e.g. neural network (Brown, Barrington-Leigh, & Brown, 2012), entropy and clustering (Zhang et al., 2011)). We adopted a very detailed grid of 768 linear regression functions. Such a detailed grid is needed to properly model non-linear daily consumption variations as well as to differentiate the behavior between seasons. For example, while in (Taşpınar, Celebi, & Tutkun, 2013) winter and summer seasons were considered we further distinguish spring and autumn that turned out to have a specific behavior. This high level of detail requires a quite large training set (i.e. one year in our tests) since the samples are partitioned through the different regression functions. We recall that a one-year training set is a standard requirement for many approaches (Brown et al., 2012). Another distinguishing feature of our model is the advanced solution for modeling user-behavior related to heaters and coolers turning off/on. Previous works in the literature (Büyükalaca et al., 2001) typically exploit a sharp threshold upon the daily average temperature. As to the anomaly detection technique most of the works in the literature (Brown et al., 2012; Chou & Telaga, 2014) defines anomalous consumptions those that are above or below two/three times the standard deviations of the consumption distribution assuming a normal distribution for residuals. However, such concepts are not deeply discussed and it is not clear how the thresholds are chosen and if the normal distribution assumption is respected. Finally, while some approaches continuously update the model exploiting techniques such as adaptative neural network (Chou & Telaga, 2014), our training approach is static. Behavior changes can be captured by periodically training the system on a sliding-window of samples. As shown in Section 4.2 this is not an issue from the efficiency point of view. As to the effectiveness point of view, on the one hand it could determine some delay in adapting the model to the new user behavior, on the other hand it reduces the risk of confusing a behavioral change with an anomaly.

We finally emphasize that similar studies have been carried out for estimating electricity consumption too (Azadeh & Faiz, 2011; Chou & Telaga, 2014; Li, Chang, Chen, & Chen, 2012). Although electricity determines different load profiles (for example electricity load is high in summer hot days due to cooling systems) most of the considerations and assumptions made for natural gas consumption forecast are valid for electricity too and the proposed forecasting techniques and the classification discussed above remain valid.

6. Conclusions and future works

In this paper we addressed the problem of short term natural gas consumption forecast aimed at supporting anomaly detection. We proposed and compared two original techniques both reaching an adequate accuracy level. We are now working towards extending the current results in the following directions:

- We are evaluating which environmental variables can lead to more effective forecast. For example some recent studies has shown that solar radiation (Soldo, Potočnik, Šimunović, Šarić, & Govekar, 2014) impacts on the consumption and can be usefully included in a forecast model. Others authors suggest further weather factors such as relative humidity and wind speed (Mirasgedis et al., 2006) but they are never been tested in a scenario similar to ours in terms of geographical zone and forecasting horizon, resolution and goal.
- We are also considering how to address the cold start problem(Lam, Vu, Le, & Duong, 2008). Our models rely on one year of historical values necessary to cover different seasons. Although network topology is quite stable if a local network is changed or a new one is added no forecast would be provided for a considerable period. This problem can be partially solved if the system can infer if there is another local network with a behavior similar to the changed/new one. We will study how behavioral similarity can be detected and how, according to such similarity, a model for known local network can be adapted to the changed/new one.
- Providing a simple forecast maybe not sufficient to support control room operators. As stated in the introduction, the main goal for HERA is to create an expert system that suggests to the operators with a low experience the right choice and that let the most skilled operators to make better and quicker decisions. Such general goals can be implemented through several functionalities each determining different research issues:
  A. Define KPIs that measure *forecast reliability* based on the analysis of the model variables (e.g. very high/low temperatures, large range of temperatures within the same days, number of available samples).
  B. Define patterns of consumption behavior to understand which of them are more probably determined by an anomaly. For example a smooth, progressive drift from the forecasted values could have a different interpretation with respect to a sharp peek of consumption. We recall that, in absence of a user-knowledge, it is hard to distinguish outliers from anomalies The first ones are defined in statistic terms as rare consumption occurrences (see Definition 2), while the second ones denotes system failures.
  C. Early-detect of anomalous consumption. Based on the existence of consumption patterns the goal is to identify an anomaly before it actually takes place. This can be done an-

alyzing the sequence of states that determines the pattern (for example by identifying the presence of a sub-pattern).

D. Suggest the best reaction to an anomalous consumption.

The proposed functionalities require a quite broad range of techniques and methodologies coming from Business Intelligence (A), Data Mining (B,C), Situational Awareness (A,B,C) and Expert Systems (B,C,D). As to the solution based on Expert Systems we are planning to exploit a rule-based approach to code the knowledge of skilled operators and to transfer it to less-skilled ones. One of the main problem we envision here is the so-called the *knowledge acquisition bottleneck* (Jackson, 1999). This refers to the difficulties computer specialists often encounter when extracting knowledge from domain experts to be transferred into expert system applications (Nilsson, Van Laere, Ziemke, & Edlund, 2008).

References

Aras, H., & Aras, N. (2004). Forecasting residential natural gas demand. *Energy Sources, 26*(5), 463–472.

Azadeh, A., & Faiz, Z. (2011). A meta-heuristic framework for forecasting household electricity consumption. *Applied Soft Computing, 11*(1), 614–620.

Behrouznia, A., Saberi, M., Azadeh, A., Asadzadeh, S., & Pazhoheshfar, P. (2010). An adaptive network based fuzzy inference system-fuzzy data envelopment analysis for gas consumption forecasting and analysis: the case of south america. In *Intelligent and advanced systems (icias), 2010 international conference on* (pp. 1–6). IEEE.

Brabec, M., Konár, O., Pelikán, E., & Malỳ, M. (2008). A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers. *International Journal of Forecasting, 24*(4), 659–678.

Brown, M., Barrington-Leigh, C., & Brown, Z. (2012). Kernel regression for real–time building energy analysis. *Journal of Building Performance Simulation, 5*(4), 263–276.

Büyükalaca, O., Bulut, H., & Yılmaz, T. (2001). Analysis of variable-base heating and cooling degree-days for turkey. *Applied Energy, 69*(4), 269–283.

Cavallo, A. J. (2004). Hubbert's petroleum production model: an evaluation and implications for world oil production forecasts. *Natural Resources Research, 13*(4), 211–221.

Chou, J.-S., & Telaga, A. S. (2014). Real-time detection of anomalous power consumption. *Renewable and Sustainable Energy Reviews, 33*, 400–411.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83*(403), pp.596–610.

Cosmo Consortium (2015). Consortium for small-scale modeling. http://www.cosmo-model.org/.

Dombaycı, Ö. A. (2010). The prediction of heating energy consumption in a model house by using artificial neural networks in denizli–turkey. *Advances in Engineering Software, 41*(2), 141–147.

Gelo, T. (2006). Econometric modelling of gas demand. *Ekonomski pregled, 57*(1-2), 80–96.

Gorucu, F. (2004a). Artificial neural network modeling for forecasting gas consumption. *Energy Sources, 26*(3), 299–307.

Gorucu, F. (2004b). Evaluation and forecasting of gas consumption by statistical analysis. *Energy Sources, 26*(3), 267–276.

Hubbert, M. K. (1949). Energy from fossil fuels. *Science, 109*(2823), 103–109.

Jackson, P. (1999). *Introduction to Expert Systems* (3rd ed.). Addison-Wesley.

Katipamula, S., & Brambley, M. R. (2005). Review article: methods for fault detection, diagnostics, and prognostics for building systemsa review, part i. *HVAC&R Research, 11*(1), 3–25.

Kizilaslan, R., & Karlik, B. (2009). Combination of neural networks forecasters for monthly natural gas consumption prediction. *Neural Network World, 19*(2), 191–199.

Lam, X. N., Vu, T., Le, T. D., & Duong, A. D. (2008). Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on ubiquitous information management and communication* (pp. 208–211). ACM.

Li, D.-C., Chang, C.-J., Chen, C.-C., & Chen, W.-C. (2012). Forecasting short-term electricity consumption using the adaptive grey-based approachan asian case. *Omega, 40*(6), 767–773.

Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*(318), 399–402.

Liu, F., Jiang, H., Lee, Y. M., Snowdon, J., & Bobker, M. (2011). Statistical modeling for anomaly detection, forecasting and root cause analysis of energy consumption for a portfolio of buildings. *12th international conference of the international building performance simulation association (ibpsa) building simulation*.

Loader, C. (2006). *Local regression and likelihood*. Springer Science & Business Media.

Ma, H., & Wu, Y. (2009). Grey predictive on natural gas consumption and production in china. In *Web mining and web-based application, 2009. wmwa'09. second pacific-asia conference on* (pp. 91–94). IEEE.

Mirasgedis, S., Sarafidis, Y., Georgopoulou, E., Lalas, D., Moschovits, M., Karagiannis, F., & Papakonstantinou, D. (2006). Models for mid-term electricity demand forecasting incorporating weather influences. *Energy, 31*(2), 208–227.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*: 821. John Wiley & Sons.

Nashawi, I. S., Malallah, A., & Al-Bisharah, M. (2010). Forecasting world crude oil production using multicyclic hubbert model. *Energy & Fuels, 24*(3), 1788–1800.

Nest-Labs (2015). Energy savings from nest. https://nest.com/downloads/press/documents/efficiency-simulation-white-paper.pdf.

Nilsson, M., Van Laere, J., Ziemke, T., & Edlund, J. (2008). Extracting rules from expert operators to support situation awareness in maritime surveillance. In *Information fusion, 2008 11th international conference on* (pp. 1–8). IEEE.

Rodger, J. A. (2014). A fuzzy nearest neighbor neural network statistical model for predicting demand for natural gas and energy cost savings in public buildings. *Expert Systems with Applications, 41*(4), 1813–1829.

Rousseeuw, P., & Leroy, A. (2005). Robust regression and outlier detection. *Wiley Series in Probability and Statistics*. Wiley.

Sánchez-Úbeda, E. F., & Berzosa, A. (2007). Modeling and forecasting industrial end-use natural gas consumption. *Energy Economics, 29*(4), 710–742.

Smith, P., Husein, S., & Leonard, D. T. (1996). Forecasting short term regional gas demand using an expert system. *Expert Systems with Applications, 10*(2), 265–273.

Soldo, B. (2012). Forecasting natural gas consumption. *Applied Energy, 92*, 26–37.

Soldo, B., Potočnik, P., Šimunović, G., Šarić, T., & Govekar, E. (2014). Improving the residential natural gas consumption forecasting models by using solar radiation. *Energy and Buildings, 69*, 498–506.

Taşpınar, F., Celebi, N., & Tutkun, N. (2013). Forecasting of daily natural gas consumption on regional basis in turkey using various computational methods. *Energy and Buildings, 56*, 23–31.

Zhang, Y., Chen, W., & Black, J. (2011). Anomaly detection in premise energy consumption data. In *Power and energy society general meeting, 2011 ieee* (pp. 1–8). IEEE.